[*International Conference on Machine Translation of Languages and Applied Language Analysis*, National Physical Laboratory, Teddington, UK, 5-8 September 1961]

INTRODUCTION TO AN AUTOMATIC ENGLISH SYNTAX (BY FRAGMENTATION)

by

M. CORBE

(UNESCO, Language Division)

and

R. TABORY

(Group for New scientific Studies, Cie. IBM-FRANCE)

PREFACE

English-to-French Automatic Translation Studies
(ETUDES SUR LA TRADUCTION AUTOMATIQUE D'ANGLAIS EN FRANCAIS - ETAAF)

This paper is Syntax Study No.1 of the above series of papers. All ETAAF papers are intended to contribute to the programming of an automatic English-to-French translation system. The texts to be processed are of a scientific, technical and administrative nature. The program as contemplated strives for exactness, intelligibility and elegance of translation, as well as for simplicity, ease of expansion and economy of analytical tools.

These studies were initiated under the impetus of the *Association pour l'etude et le developpement de la traduction automatique et de la linguistique appliquee* (ATALA, 20 rue de la Baume, Paris 8e).

In their present form they are the result of a joint effort by a group of members of this Association, a certain number of independent researchers and the group for New Scientific Studies (Cie. IBM-FRANCE, 5 Place Vendome, Paris ler).

Linguistic and operational research is carried out under Michael CORBE, while programming studies are conducted under Robert TABORY.

INTRODUCTION

Scope and Organization of Study

SYNTACTICALLY speaking, any "translation" - whether it involves two natural languages, one natural and one artificial language, or simply a text to be interpreted or summarized in its original language - raises three types of problems:

A. - Definition and recognition of (a) the syntactic patterns of the original, and (b) the relationships that make them into a formally plausible whole (input syntax);

B. - Definition and recognition of equivalent, if not entirely identical, structures and relationships in the target language (Output syntax);

C. - Correlation of the two sets of phenomena (input-to-output transfer procedures).

In the present series, these three groups of problems will be discussed in turn. In addition, in view of the above definition of the term "translation", Stage B will be subdivided into:

Bl. - Automatic syntax of English abstracts;

B2. - Automatic syntax of French abstracts;

B3. - Automatic syntax of full French translations of original English texts.

As far as English interpretation of English originals is concerned, we think it unnecessary to devote a special subsection to this problem in view of the fact that it is dealt with both in transformation and lattice theories.

Within this framework, *Sub-stage Bl i*s expected to provide a quick and economical way of verifying the validity of patterns defined in the course of *Stage A. Sub-stage B2* will be used for testing the validity of our translation procedures involving two semantically different languages, on the basis of a simplified syntax. Finally it is hoped that *Sub-stages Bl, B2 and B3* considered as a whole will make it possible to evolve standardized analytical tools, equipment and recognition procedures for:

(1) Automatic production of English "digests";

(11) Their automatic translation in French;

(111) *In extenso* translation of any original text found to be important enough on the basis of its pre-digested version.

From the outset, our work was influenced by Victor H. Yngve's ideas concerning the need for an exhaustive syntactic analysis of both languages involved, and A. Sestier's perspicacious and convincing interpretation of same (1) (2). In the course of our studies we further discovered that in some respects our way of thinking was very much akin to the method of "clause and phrase" openers advocated by Franz L. Alt (3), as well as to the as yet unpublished work of I.A. Mel'cuk (USSR) concerning a Russian "syntactic dictionary" which seems to resolve itself into a series of stable syntactic patterns. Our general way of tackling the problem of MT is fairly close to the one propounded by the Rand Corp. and Ramos-Wooldridge teams in Los Angeles: while as a matter of principle we work only on the basis of actual texts, we do not hesitate to check our conclusions against sensible counter-examples generated *ad hoc.*

As far as our equipment requirements are concerned, they are definitely oriented towards a very large memory provided with a relatively rudimentary logic limited to a series of essentially identical look-ups. A buffer device allowing for repeated scannings of a sentence or even a whole paragraph would also be useful. Access time is of no primary con-

cern to us for we believe that there should exist an *optimum* relationship between this particular factor and the cost of production and of operation of the machine. The possibility of building such a specialized device was indicated about thirty years ago by G.B. Artsrouni to whom we are very much indebted in this respect (4). We are now viewing with the same interest an essentially similar machine (except for the fact that it is photoscopic instead of electromechanic) that has been developed by Dr. Gilbert W. KING for the US. Air Force (AUTOMATIC TRANSLATOR MARK II) (5). Our linguistic and operational analysis owes a great deal to Dr. Carl Mayer's approach to the "problem of social stratification" as taught by Dr. Mayer himself a few years ago at the New School for Social Research in New York.

Within these limits, however, we intend to remain as open-minded and eclectic as possible.

All working hypotheses and procedures outlined in the present paper are concerned with *Stage A* and *Sub-stage Bl* as defined above. They are based on the results of a preliminary examination of a limited English corpus. These results, as well as our first working definitions, are reproduced in their original form in the *Appendix* which the reader may well wish to consult before going over to the sections below.

A. INPUT SYNTAX

1. CONVENTIONAL SYNTACTIC ANALYSIS

(a) Manual definition and recognition of syntactic patterns. - In classical syntax, words are defined in terms of the parts of speech they are made to belong to in the dictionary. When placed in a specific situation they are assigned syntactic functions such as the "subject", the "predicate", the "object" etc. In this guise, they form syntactic units defined as "clauses" and "phrases" which, in turn, are combined into "sentences".

The recognition of these elements bolls down to the need for:

-an unambiguous identification of all the parts of speech contained in a sentence;

-an unambiguous identification of their syntactic function in any given situation;

-a valid recognition of the situation itself, that is of the upper and lower limit of each of the syntactic units Involved.

In natural languages, the elements to be recognized are often mutually determined, i.e. the length of a given unit depends on the syntactic function of each of its components, and *vice versa.* In manual syntax, this difficulty can be overcome to the extent that the student of the language has the possibility of switching back and forth from one recognition level to the other. Thus, he is able constantly to adjust his findings until all the units he has identified fall without overlapping into their proper place within the sentence.

(b) Manual definition and recognition of inter-pattern relationships. - These relationships are defined in terms of a hierarchy established between different kinds of syntactic units. Their recognition is the last and necessary check in syntactic analysis because it is perfectly possible to dissect a sentence into individually plausible but mutually incompatible units, such as: THE MOMENT HE SIGHTED/HER MOTHER/HE WENT BERSERK./ instead of: THE MOMENT/HE SIGHTED HER MOTHER/ HE WENT BERSERK./, and such errors must be immediately spotted and corrected.

It is obvious, however, that this check can only be applied after all the problems listed under (a) have given rise to suitable working hypotheses. The usual procedure then consists in reconstructing the original sentence in accordance with certain combinatory "do's" and "don't's" provided for by the hierarchical rules. In other words the sentence is rebuilt around a particular syntactic unit such as a subject or a predicate clause placed at the top of the hierarchy by convention.

(c) Attempts at mechanizing conventional input syntax. - Several attempts have been made at mechanizing recognition procedures on the basis of conventional (if somewhat modified) pattern definitions. In view of the lack of intuition in machines, the "back and forth" syntactic scanning had to be replaced by a word-for-word analysis supplemented with *ad hoc* sub-routines designed to take care of syntactic ambiguities. In this connection the following figures were given:

Group A, Russian-to-English: about 36,000 instructions needed for handling a declarative sentence of 10 to 15 occurrences. Of these - 17,000 instructions for syntactic analysis alone. *

* This figure was given to us by several American sources during our 1959 trip to the U.S.

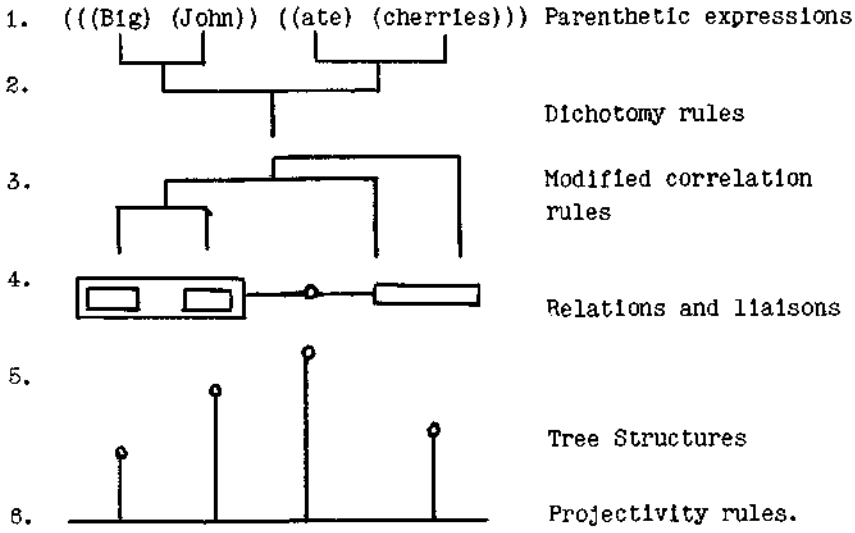
Group B, French-to-Russian: about 50,000 instructions for a sentence of approximately 8 to 10 occurrences, and the proportion of syntactic sub-routines is probably the same. (6).

Admittedly, both of these groups are working on a "95 per cent" basis. That means that while they claim to be able to identify 95 per cent or so of occurrences, the proportion of sentences made intelligible is consider-ably less (7). Any further reduction of the "5 per cent recognition gap" would entail a disproportionately high number of additional instructions.

In English, this very serious "economy vs. efficiency" problem is further compounded by the exceptionally high degree of external and mixed ambiguity, the length of individual sentences and the elliptic propensities of the language.

(d) Special difficulties in mechanizing conventional English syntax. - These difficulties can be illustrated as follows:

(1) *Syntactic ambiguity.* - Let us consider the sentence BIG JOHN ATE CHERRIES.
It contains no part-of-speech ambiguities and thus can be analyzed in accor-dance with the most elementary linear procedures that lead up to any of the following representations:



But let us take

THAT MAN LIKES SHIPS(.)

instead, and we immediately find ourselves in trouble. While at the last recognition level there is no difficulty whatsoever in representing this sequence in exactly the same manner as the previous one, its *a priori* part-of-speech analysis cannot but yield the following results:

THAT (a) Demonst. pronoun, 3rd pers. sing., neuter;
(b) Demonst. adj., 3rd pers. sing., gender determined by that of the noun modified;
(c) Subordinating conjunction

MAN (a) Adj.;
(b) Noun, animate masc. sing.;
(c) Verb, trans.: infinitive or imperative (sing, or plur.); present indicative (W - 3rd pers.) or present subjunctive (W pers.).

LIKES (a) Noun, animate or inanimate; masc., fem. or neuter; plur.;
(b) Verb, trans, or intrans.; present indicative 3rd pers. sing.

SHIPS (a) Noun, inanimate; fem. or neuter depending on usage; plur.;
(b) Verb, trans, or intrans.: present indicative, 3rd pers. sing.

(.) (a) Full stop;
(b) End of abbreviation;
(c) Decimal sign inside a fraction;
(d) Decimal sign introducing a fraction;
(e) Beginning of a suspension mark.

In other words, three of the five occurrences present are potential SUB or VER, while the two remaining ones signal other types of ambiguity.

Under these circumstances the construction of any of the above mentioned graphs becomes impossible without introducing a large number of external criteria and/or very elaborate checking and cross-checking procedures.

This syntactically 100 per cent ambiguous sentence only serves to dramatize the following point brought out by our studies to date: The average degree of syntactic ambiguity in an English corpus of approximately

1,000 occurrences can actually fluctuate between some 40 per cent and slightly more than 70 per cent. That is to say that some of the authors studied do not hesitate to use sentences with an 85 per cent syntactic ambiguity rate.

Another difficulty connected with this particular type of ambiguity is the uncertainty that prevails as to the length and the location of the context to be considered for determining the syntactic function of a particular word in a particular situation. Let us consider the word HER, for example, which may stand for:

-the objective case of SHE

-the possessive case of SHE

-an adjective related to a fem. antecedent.

Any choice between these three functions must take into account the existence of the following and, to be sure, many other combinations:

-HE GAVE HER AWAY.

-HE GAVE HER MONEY.

-HE GAVE HER MONEY AWAY.

-HE GAVE HER MONEY AWAY FROM HOME.

-HE SAW HER MOTHER AWAY FROM HOME.

-HE GAVE HER .5 PER CENT OF HIS SHARES.

-HE LIKED HER THE MOMENT HE SAW HER.

-HE GAVE HER THE AMOUNT HE OWED HER.

-THE MOMENT HE SAW HER MOTHER STOPPED CRYING, etc, etc.

Here again, a fully satisfactory syntactic analysis of a single occurrence based on conventional definitions would require a very large if not unmanageable number of instructions.

To obviate the difficulty one may contemplate cutting down the number of part-of-speech indices in the dictionary and thus creating a kind of syntactic microglossary.

However, such a procedure not only would restrict the field of application of any MT system but would generate serious analytical errors as well. The following examples relating to the word IN should suffice to illustrate this point:

- HE CAME IN.
- THE CAT IS IN THE BAG.
- THE IN TRAIN IS LATE.
- HE WENT OUT TO IN THE HAY.

Failure to provide for appropriate part-of speech indices would make correct translation of these sentences more or less impossible. (And think of the beauty that appears on match boxes sold at all US. Army commissaries: RE-UP! ARMY! NOW!). Conversely, a sufficiently refined and judicious syntactic indexing of individual dictionary items may go a long way even toward removing some of the semantic ambiguities.

(ii) *Overall length of sentences.* - In strongly inflected and well articulated languages, such as Russian, the very length of a sentence can facilitate the task of resolving syntactic ambiguities because the number of the latter rises at a slower rate than the total number of occurrences. In English, however, quite the contrary is the case.

Thus, our initial sentence
THAT MAN LIKES SHIPS(.)
can be legitimately expanded into
THAT SHIVERING CAVE MAN LIKES WATCHING SPACE SHIPS THAT SAIL HIGH UP IN THE SKIES(.)

While the sentence is still 100 per cent ambiguous the ratio of potential verbs has risen from 3/5 to 11/16.

Absurd as it sounds, this particular construction is by no means outlandish. What is more, *Table 1* of the *Appendix* indicates that sentences containing 20 occurrences or less cover less than 27 per cent of a total of 1,000, while more than one half of all the corpus is concentrated in sentences of between 25 and sixty occurrences (with a more than proportionately increased degree of *a priori* syntactic confusion).*

(iii)*Elliptic patterns.* - Here again, a few examples should suffice to illustrate this point:

- I THINK YOU ARE ILL
- THE RING I GAVE YOU
- HE WENT IN.

* See ANNEX, *Table V below*

The danger of the occurrence of such patterns is ever present in English and should be constantly kept in mind. The solution of this problem would require new and considerable additions to the store of necessary instructions.

(e) **Conclusion:-** It appears then that if we were so much as to try to mechanize English syntax on the basis of conventional definitions we would inevitably and very quickly run into tens of thousands of instructions for syntactic analysis alone, without any guarantee of completeness. To be economical, such a system would have to rely on electronic wonders of speed and efficiency produced at a much lower cost than the existing ones, which is rather unrealistic.

It is for this reason that we believe that attempts at mechanizing conventional input syntax should be abandoned, and that a new definition of this linguistic discipline should be sought for machine translation purposes. Hence the distinction we are trying to establish between the "conventional" (whether manual or mechanized) and "automatic" input syntax.

2. AUTOMATIC INPUT SYNTAX

(a) **Mnemonics vs. rationalization**. - While conventional syntax is taught in every high school, little of it is actually used in our everyday handling of languages. The authors, for their part, would feel considerable embarrassment if they were to perform here and now as complete a syntactic analysis of their own text as they would have been able to in their fifth grade. Nevertheless, they believe they are expressing their thoughts correctly and articulately, if not with all the elegance desired.

The same thing applies to young children who have never heard of conventional syntax, and yet are using the vocabulary available to them in a perfectly commendable manner.

These preliminary remarks are corroborated by the experience of one of the authors of the present paper, who has been associated for years with the linguistic services of several international organizations dealing with a dozen or so different languages.

It appears to him that a translator uses his conventional syntactic knowledge only when confronted with a language he does not know well. In such a case, he analyzes the original text sentence by sentence, clause by clause and even occurrence by occurrence, which is a protracted and rather

uneconomical process, yielding more often than not linguistically unsatisfactory results.

If this is so, it is because the linguistic expression of the original conveys no meaning to the translator who thus finds it necessary to reconstruct the logical rather than the linguistic patterns before him.

Conversely, a good translator endowed with a complete mastery of the input language does not "think." He proceeds quickly and automatically by simply taking in and digesting groups of occurrences lodged in his memory as "legitimate", while rejecting those that are not. Thus, the following Spanish sentence can be grasped and translated without much analytical ado by means of the following grouping:

DURANTE ESTE DEBATE
SOBRE ESTE INFORME
DEL DIRECTOR GENERAL
COMMA
HEMOS ESCUCHADO VOCES
DE MUCHOS PUEBLOS
FULL STOP,

while such groups as:

DURANTE ESTE
DEBATE SOBRE ESTE
INFORME DEL
DIRECTOR
GENERAL, HEMOS
ESCUCHADO VOCES DE
MUCHOS PUEBLOS (.)

must be rejected as illegitimate.

As can be seen, all legitimate groupings are purely automatic. The only "logic" involved concerns the recognition of prepositions, verbs and punctuation marks, as well as the decision to slice the sentence on the basis of these particular diacritics. The groups identified by the above mentioned means are recognized as such and then replaced by similar prestored units in the output language.

It is thus legitimate to surmise that while conventional syntax is essential for understanding the relationships between thought and language, our actual handling of the latter is based on short-cuts, that is habit and memory. This situation is best reflected in the motto familiar to all international simultaneous interpreters, "Once you start thinking about what you are saying, you are lost."

The crux of the matter lies apparently in the fact that, at the beginning of our speaking careers, we must be memorizing not only words but also whole sentences. This process is particularly noticeable in bilingual children who keep switching from one language to the other without even knowing it and without being able to unify their way of expression, simply because they have memorized certain things in one language and not in the other.

After the number of concepts and their actual combinations have grown too large to be stored in full in a child's mind, abstract sentence patterns are developed which could be compared to beach buckets serving for the production of identically shaped "sentence cakes" whatever the actual word material used.

At a still later stage full-sentence patterns are replaced by truncated patterns that are then combined, the way play cubes would be, into much longer sentences according to need.

It is usually at this stage that conventional syntax intervenes. It is our contention, however, that after school most of us forget almost everything about that particular kind of syntax and continue happily chattering away much as before.

Reverting to the English language, we may state that if we have no difficulty in understanding our initial sentence
THAT MAN LIKES SHIPS(.)
it is not because of any elaborate analysis on our part but simply because we *remember* all possible part-of-speech combinations it can give rise to, as well as the fact that only one of them, namely:
DEMONST. PRON., SING., MASC. of a certain type
NOUN, ANIMATE, SING., MASC. " " " "
VERB, TRANS., PRES. INDIC., 3rd PERS. SING.
NOUN, INANIMATE, FEM. or NEUTER, PLUR.
FULL STOP
is legitimate, while all others are not.

In other instances we even know that according to conventional syntax the expression

I IS SICK

is wrong, but we also know that it is frequently used by certain groups of population, and we remember it as such for future reference.

The aforementioned considerations lead us to the tentative conclusion that an automatic (as opposed to conventional) syntax should consist in the following:

- Definition and storing of an exhaustively large number of "legitimate" part-of-speech combinations both in their ambiguous and fully determined form;
- Recognition and retrieval procedures;
- Operational rules automatically indicating the possibility, for two or more fully determined patterns, of entering into even larger combinations, and defining the nature of such combinations.

To be universally valid, such a syntax should be designed to cover any type of sentence of arbitrary length (see *Table I, Appendix).* As such, it would require the redefinition of a certain number of conventional syntactic concepts.

(b) **Automatic input pattern definitions.** - The main body of dictionary - or "static" definitions is listed in the Appendix. The following "operational" definitions should be added for analytical purposes:

- *Fragment:* This is a fundamental concept. It is defined as a group of parts of speech that cannot be subdivided any further. It is long enough to convey sufficient syntactic information, and yet short enough to occur repeatedly in any type of sentence. It is expressed in abstract form and recognizable a priori. Only fragments at the outset recorded as such are deemed to be legitimate, while all other linear combinations of occurrences are not.

While in some respects a fragment resembles the Altian "clause" or "phrase", in some others it is closer to Molosnaja's and Mel'chuk's "configurations" or even to Martinet's "autonomous syntagmatic units". From still other points of view it is, however, entirely different from all of these and can take the following "barbarous" form: He TOLD ME HE was ill. A more refined definition of this concept will be given after the concept

of "separator" is explained. The importance of these differences will be brought out when the subject of elliptic structures is discussed.

Our experience to date indicates that any English sentence can be broken down into repeatedly occurring fragments of two to six (possibly seven) occurrences. The most frequent ones contain, however, between three and five occurrences.

The hypotheses to be tested in this connection are:

- There is only one set of legitimate fragments to represent a sentence;
- There is one and only one fully determined fragment for each ambiguous one, while the contrary is not necessarily true;
- The number of legitimate fragments sufficient completely to circumscribe a syntax is electronically "manageable", that is does not run into tens of thousands. (The answer to this question can, however, be controlled to a certain extent because the number of ambiguous fragments depends on the degree of refinement of syntactic indexing to be established empirically. Thus, depending on how IN, OF and FOR are indexed, the fragments
  IN CLASSICAL LOGIC
  OF CONVENTIONAL LINGUISTICS
  FOR YOUR INFORMATION
  a posteriori resolve themselves into the single formula PRE ADJ SUB tagged with certain additional characteristics can be recorded either as a single ambiguous fragment or as two or even three different possibilities. It follows that in the course of our studies *optimum* solutions will be sought between the degree of determination and the number of fragments).

- *Separator (SE):* A separator is tagged in the dictionary as a boundary word liable to introduce a fragment. A separator is said to be *absolute* (SEA) if the word belongs to an *a priori* fully determined part of speech (A_i), and *conditional* (SEC) if the part of speech (X_j) is ambiguous.

Generally speaking, the separator quality is attached to certain actual or potential verbs, prepositions, conjunctions, relative pronouns and punctuation marks. By convention, no SEA is permitted to occupy within a fragment a position other than the first, except in cases when it follows another SEA of the same type (Ex.: HE SHOULD HAVE COME). In such a case it loses its privileged position. Similarly, if a SEC is found within the

boundaries of a legitimate fragment it loses its separator value and has to be interpreted as a non-separating word. Thus a fragment can now be defined as a *sequence of parts of speech introduced by an effective SE and ending immediately before the next effective SE, independently of whether these are absolute or only conditional at the outset.*

Some of the SEC are so ambiguous that they have to be considered as separate entities to be unscrambled during the third or "relationship" scanning. They are provisionally listed as (.), (,), AND, THAT, WHICH and WHO, and there is no doubt that several others will have to be added in the course of our subsequent studies.

- *Syntactic function:* The ability of a part of speech to enter into fragments of a given type and length.

- *Part of speech:* The definitions listed in the *Appendix* were used for purely illustrative purposes and need further refinement for automatic syntactic analysis. Each of the part-of-speech categories alluded to in the *Appendix* must be further subdivided on the basis of criteria which may be syntactic, morphological or even semantic according to need. Thus, the VER category is now made to contain sub-categories based on conjugation deficiencies, semantic contents (movement, feeling, order, etc.), prepositional possibilities, and the like.

The listing and expansion of all relevant categories and sub-categories is now under way, and we expect to wind up with less than a few hundred different parts of speech. It should be even possible gradually to shorten this list on the basis of experience by eliminating characteristics that would prove useless in actual operation.

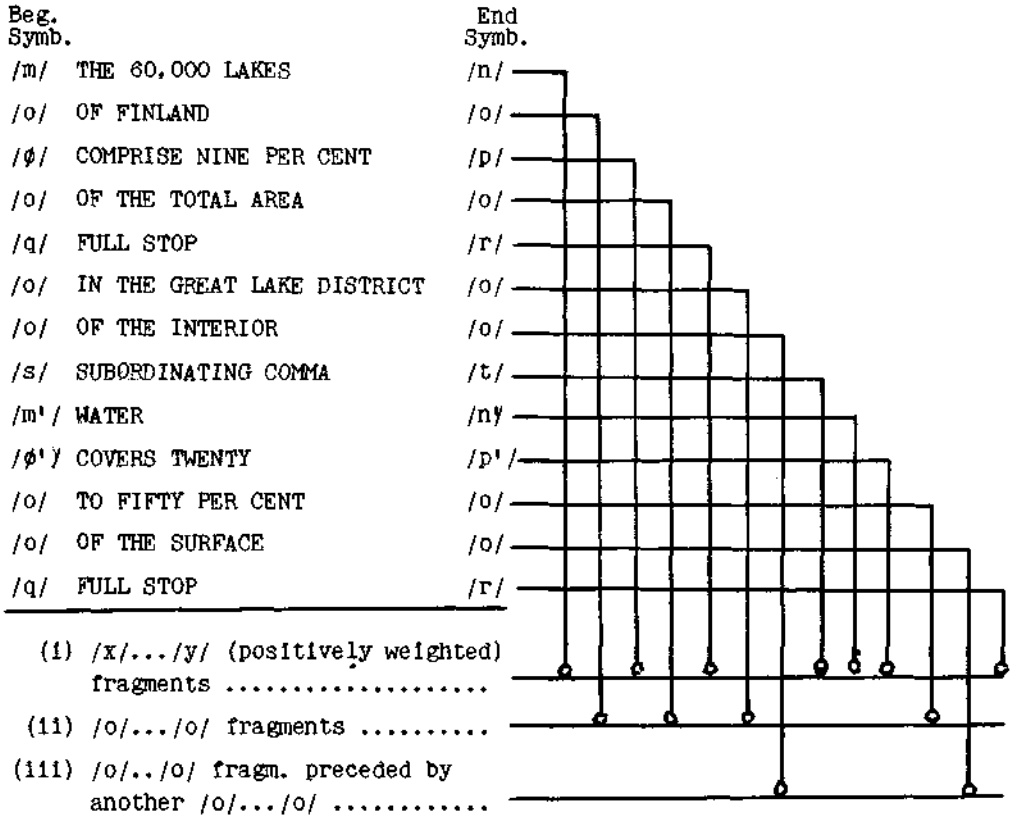
The total number of indices to accompany any dictionary item (word) will have in each case to be decided upon empirically. We do not expect the total volume of dictionary information to exceed 250 binary digits per word.

(c) **Automatic input relationship definitions.** - Interfragment relationships are defined in terms of symbols preceding and following each fully determined fragment. Some of them have zero value indicating that their relationship with the rest of the sentence is irrelevant for analytical purposes. Others have a positive weight attached to them. Combinatory possibilities are important only between positively weighted fragments.

However, as mentioned above, certain SECs remain by definition outside the system of fragments and have thus an intrinsically ambiguous weight. The same thing applies to other ambiguous words deliberately left in isolation by virtue of our fragmentation rules. Their final weight can only be determined on the basis of their comparison with the surrounding fragments.

After this is done, combinations of any two neighboring non-zero "end"- and "beginning" symbols indicate the type of relationship linking the two fragments involved. (By that time all isolated words have also become "fragments", and all zero value fragments have been eliminated). These combinations carry instructions to be effected for the transformation of any of the positively weighted fragments involved.

The system of weights can be illustrated by the following graph wherein (.), (,) and WATER are presumed to have been weighted by appropriate means. On this graph, different weights are represented by Levels (i), (ii) and (iii):



It is obvious that only Level (i), that is: /m/.../n/ - /ø/.../p/ - /q/.../r/ - /s/.../t/ - /m'/.../n'/ - /o'/.../p'/ - /q/.../r/ is relevant for final inter-fragment relationship determination.

In this particular instance, the "end" - "beginning" symbols indicate that a VER introduced by /ø/ or /ø'/ agrees in number and/or gender with the SUB at the core of the fragment ending with /n/ or /n'/.

/r/ signals the end of a sentence. Thus, all .../r/-/x/... end-beginning pairs indicate the beginning of a new analytical cycle. The pairs /r/-/s/ show that the second sentence is introduced by a subordinated clause to be neglected in final analysis.

It also should be noted that identical "beginning" and "end" symbols can be attached to fragments of different length, thereby indicating identical relationship possibilities:

LAKES

THE LAKES

THE BEAUTIFUL LAKES

THE SIXTY THOUSAND LAKES

THE SIXTY THOUSAND MOST BEAUTIFUL LAKES

THE SIXTY THOUSAND CONSIDERABLY MORE BEAUTIFUL LAKES

All of these will have in the end the same /x/.../y/ weight.

(d) **Stored knowledge*** - All information based on the above definitions is stored in the machine memory in the following form:

- *English dictionary:* Our concern with program economy makes it essential to have a paradigmatic dictionary (See the definition of "word" in the *Appendix).* Each dictionary item is accompanied by indices reflecting its actual or potential syntactic function, which is either of the type A_i or of the type X_j. In addition, each dictionary item is provided with a symbol indicating whether it belongs to a SEA, SEC or SE-zero category.

- *Suffix table:* The distribution of all occurrences between the A_is and the X_jS is essential for the recognition of fragments. Thus, at

* Wherever possible, we are deliberately adopting the terminology coined by Victor H. Yngve in the article quoted under (1).

least in the initial stages, special means will have to be provided for the part-of-speech determination of words not recorded in the dictionary. This entails the provision of stripping routines, as well as the existence of a special table with all suffixes classified according to the parts of speech the unknown word may belong to. Thus, each suffix is implied with either an A_i or an X_j index.

- *Tables of fragments:* These tables are classified according to the length and nature of fragments (declarat., interrog., etc). They contain all ambiguous fragments found to be legitimate, and their fully determined counterparts with appropriate "beginning" and "end" symbol pairs indicating their weight /x/.../y/.

A special section indicates the weights of "isolated words", whether these weights are fully determined (/x/.../y/) or ambiguous (/w/.../z/).

- *Table of inter-fragment weight combinations:* This table is designed to eliminate all the /w/.../z/s. In certain cases it carries instructions for a subsequent modification of the original weight of an /x/.../y/ fragment.

- *Table of positively weighted "end" - "beginning" symbol combinations:* In this table all legitimate non-zero . . ./y/ - /x/... combinations are listed, together with the necessary instructions for intra-fragment transformations.

(e) **Recognition and retrieval procedures.** - In addition to the above mentioned tables, the stored knowledge comprises instructions necessary for their use, and especially for switching from one table to the next one.

It is estimated that four scanning stages will be sufficient for the execution of all the instructions. The first three stages will be concerned with pattern recognition and retrieval. The last one, with inter-fragment relationship.

Thus, if a sentence were to be represented as the following sequence of occurrences:

ø1 ø2 ø3..... øn-l øn,

the instructions would be as follows:

-First scanning: Application of traditional routines:

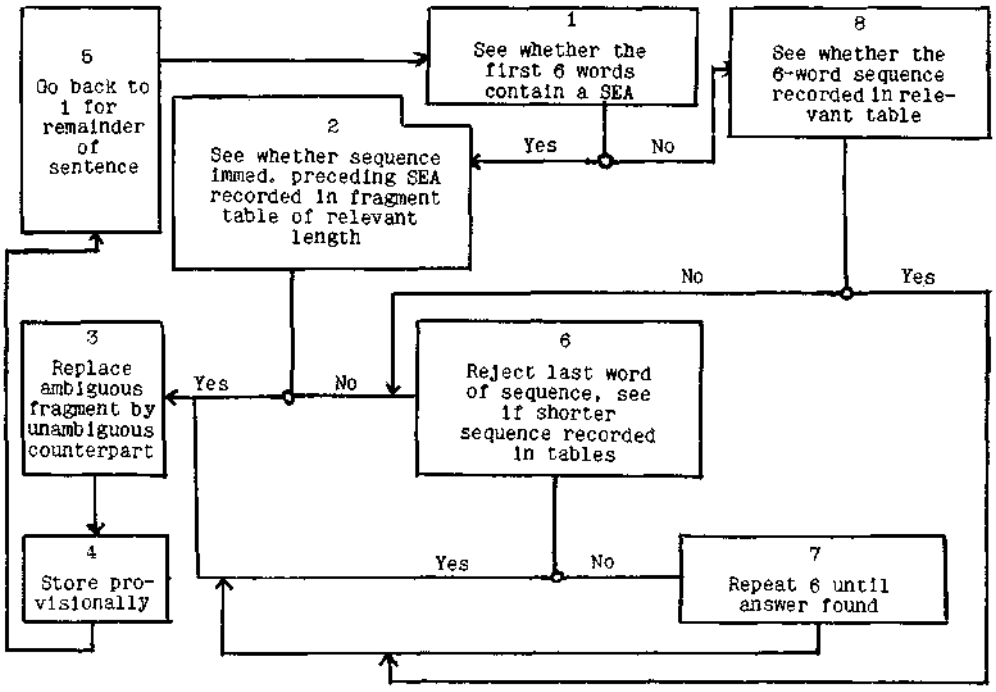
Dictionary look-up
Stripping of unknown words for the purpose of their indexing
Identification of proper nouns
Provisional storing of syntactic information thus retrieved

As a result, the original sentence would be replaced in the buffer by an abstract sequence of the following type:

A11 Xj2 A13 Xj4 A15 A16 Xj8 A1n-1 ...Xjn,
wherein ☐ would symbolize all SECs and ▭ all SEAs.

This abstract sequence would then be studied in the course of the second scanning.

-Second scanning: It deals with the dissection of the previous sequence into legitimate fragments and their replacement by fully determined ones. To this end the following procedures would be used:



If the original abstract sequence contained no "isolated words", application of the above procedures would yield a string of fully determined fragments:

/x/ A_il A_i2 A_i3 /y/
/x'/ A_i4 A_i5/ y'/
/x''/A_i6 ..../y''/, etc. These fragments could be immediately submitted to the fourth and last scanning.

However, if the results of the second scanning were:

/x/A_il A_i2 A_i3/y/ /w/X_j4/z/ /x'/ A_i5 A_i6 /y'/ ... etc.,
application of the third scanning stage would become necessary.

- *Third scanning stage:* At this stage all /w/ ... /z/s would be looked up in their environment tables and replaced by /x/ ... /y/s. As a result, X_js would be replaced by A_is. AS a corollary, the intrinsic weight of some of the original /x/ ... /y/ fragments would have to be changed on the basis of instructions contained in these tables.

In this manner, all fragments would be supplied with finality with their proper weight.

- *Fourth scanning:* At this stage, all /0/ ... /0/s would be set aside and strings of positively weighted fragments examined on the basis of "end" - "beginning" symbol tables. Instructions contained in these tables would be applied for intra-fragment transformations, if necessary, as well as for establishing the final hierarchy of all fully determined fragments.

This fourth scanning should conclude the cycle of the automatic syntactic analysis of a sentence. Theoretically it should produce a complete abstract representation of any English sentence dissected into non-overlapping fully determined fragments with univocal indication of their respective weight, mutual relationship and hierarchy. In the last analysis these results would be as significant for machine translation as those obtained by conventional syntactic procedures.

In its initial stages the system as contemplated would of course require numerous additions and/or revisions. Thus if, after the fourth scanning stage, ambiguous fragments persist in the sentence, this would mean one of the two following things:

(i) the dissection is wrong, that is the quality of SEA has wrongly been attributed to a SEC, a SEA has been neglected

or an "isolated word" not sufficiently taken into account. In such a case, all relevant fragments and tables would have to be revised.

(ii) One or more of the legitimate fragments contained in the sentence are of an unrecorded type. In that case additions would have to be made.

(f) Machine-to-man-to-machine feed-back with regard to initial definitions of fundamental concepts. - At present we are working with an English corpus which has been specially prepared for us by IBM Research Laboratories at Yorktown Heights, N.Y. (Dr. Gilbert W. King, Research Director). We are planning to use the dictionary of this corpus and to index therein, in the first place, all the separators. The corpus will then be dissected automatically. The validity of the fragments thus obtained will be tested manually, and necessary corrections will be made.

At the same time we shall seek the necessary and sufficient amount of syntactic information to be attached to each word. All relevant characteristics will be added and recorded progressively, on the basis of the preliminary machine findings. After a first minimum indexing a machine will be entrusted with the task of replacing all the sentences of the corpus by abstract sequences which will be dissected again. The categorization of words will be recast and completed manually on the basis of these results.

It is also at this stage that /w/ ... /z/ and ... /y/ - /x/ ... tables will be inserted. Further tests will be made to ensure their universal validity.

B. AUTOMATIC OUTPUT SYNTAX

1. ENGLISH DIGESTS

As mentioned before, this Sub-stage is primarily intended to test the validity of the conclusions reached as a result of *Stage A.* It is much more economical indeed for testing purposes to switch from a certain type of English syntax to another type of syntax of the same language than to undertake a complete syntactic analysis of a semantically different way of expression.

Secondly, this testing Sub-stage may bring about an immediate payoff in that it would permit producing English digests of English texts on the basis of the same dictionary, same tables and same procedures, with only

one set of additional tables inserted at the production end.

(a) **Condensing English fragments.** - The whole idea rests on the experience of the United Nations linguistic services in N.Y.

Initially all summary records of the Organization were written in the usual analytical manner, in one of the two official languages. This procedure, however, was found unsatisfactory by many delegates who complained that they were not able to recognize their own speeches.

As a result, the "blue pencil" method was adopted. Each text was considered in its original version, abridged on the basis of certain formal criteria, and only then translated into one of the working languages. The average degree of condensation requested was (and we believe still is) between 60 and 65-70 per cent.

If such a procedure were to be mechanized on the basis of fragments, each fully determined fragment would be supplemented in the tables by an abridged one which would then be printed at the output side.

According to the degree of condensation desired the previously quoted sentence would read:

- *On a literary basis:*

THE SIXTY THOUSAND LAKES
OF FINLAND
COMPRIZE NINE PER CENT
OF THE AREA
FULL STOP
IN THE DISTRICT
OF THE INTERIOR
SUBORDINATING COMMA
WATER COVERS TWENTY
TO FIFTY PER CENT
OF THE SURFACE
FULL STOP

- *On a telegraphic basis:*

SIXTY THOUSAND LAKES

FINLAND

COMPRIZE NINE PER CENT

AREA

FULL STOP

IN DISTRICT

INTERIOR

SUBORDINATING COMMA

WATER

COVERS TWENTY

FIFTY PER CENT

SURFACE

FULL STOP

- *Suppressing unnecessary original fragments:*

LAKES

COMPRIZE NINE PER CENT

FULL STOP

WATER

COVERS TWENTY

FULL STOP

The optimum degree of condensation can be decided upon during the knowledge-storing stage. It can be controlled by the semantic indexing of all dictionary items and by the form of the output fragments to be inserted in the tables. The achievement of this optimum requires no additional operational procedures during the actual analysis of a text.

(b) Conclusion. - If the hypotheses described above are verified on the basis of a set of corpora of sufficient size, it seems to us that syntax by fragmentation could become a fairly powerful analytical tool. It would fully describe the input language syntax without tying it to the output language.

As such it could serve to convert English texts into any and all equally fragmentable languages.

In addition to its simplicity this method would offer the advantage of producing better than word-for-word translations in view of the fact that output fragments would not have to follow exactly the same patterns as those in the input language.

Finally it seems to us that it would offer interesting possibilities for studying the English language itself, and probably quite a number of other languages as well.

REFERENCES

1. YNGVE, V.H., "A Framework for Syntactic Translation" in *Mechanical Translation,* 1957, **4**, No. 3. (Published at the Massachusetts Institute of Technology, Cambridge, Mass.)

2. SESTIER, A. "La Traduction Automatique de Textes Ecrits Scientifiques et Techniques d'un Langage a l'Autre" in *Ingenieurs et Techniciens,* No. 120, 121, and 122 (April, May and June 1959), Paris.

3. ALT, F.L. "Recognition of Clauses and Phrases", *National Bureau of Standards Report* No. 6895.

4. CORBE, M. "La Machine a Traduire Francais Aura Bientot Trente Ans." in *Automatisme,* 1960, **5**, No. 3 (Published by Dunod, Paris) Mr. G.B. Arstrouni died a few months after the publication of this article, leaving a considerable body of unfinished research.

5. KING, Dr. G.W. "Final Report on Computer Set AN/GSQ - 16 (XW-1)", 1959, **1**, The Photoscopic Memory System. Yorktown Heights, N.Y.

6. For a more detailed description of MT work with Russian as target language, see: EVREINOV, E.V., and KOSAREV, Ju. G: OB EFFEKTIVNOSTI ISPOL'ZOVANIJA UNIVERSAL'NYH VYCISLITEL'NYH MASIN DLJA CELEJ PEREVODA in *Doklady na konferencii po masinnomu perevodu, obrabotke informacii i avtomatices komu cteniju, Moscow, 1961.*

7. DELAVENAY, E.: "Introduction to Machine Translation". Appendix, Thames and Hudson, London, 1960.

APPENDIX

RESULTS OF A PRELIMINARY EXAMINATION OF A LIMITED ENGLISH CORPUS

To get at least some idea of the nature of the difficulties involved in the definition and recognition of English syntactic patterns, we have made a preliminary study of an English corpus selected entirely at random. (See text at end of Appendix).

A. Definitions

To that end the following definitions were adopted:

- *Letter:* A letter of the English alphabet without distinction between capitals and lower case letters.

- *Word:* A sequence of one or more letters, bordering on its left and right on blank spaces and distinct from sequences of a different composition. Thus the morphological forms *sister, sisters, sister's* and *sisters'* are considered to be different English words while *mothers* (Noun) and *mothers* (Verb) are listed under a single dictionary item. Punctuation marks are taken to be words.

- *Occurrence:* A word, each time it occurs in the text.

- *Sentence:* A sequence of occurrences bordering on:
  - blank spaces extending to its left and right all over the length of a line, or
  - a blank space on its left, and a full stop, an exclamation sign, a question mark , a suspension mark or a full stop and a quotation mark on its right, or else
  - one of these punctuation marks on both sides of the sequence.

- *Fully determined part of speech:* A group of words always fulfilling the same syntactic function and recognizable as such *a priori.* As a first approximation these parts of speech are defined in accordance with the indications given by Webster's Collegiate Dictionary:

ADJECTIVE	(ADJ)
ADVERB	(ADV)
DEFINITE ARTICLE	(ARD)
INDEFINITE ARTICLE	(ARI)

CONJUNCTION	(CNJ)
INTERJECTION	(INT)
PROPER NOUN	(NOP)
PREPOSITION	(PRE)
PRONOUN	(PRO)
NOUN	(SUB)
VERB	(VER)
NUMBER	(CHI)
FORMULA	(FOR)
PUNCTUATION MARK	(SIP)

- *Externally ambiguous part of speech:* A group of words belonging to two or more mutually exclusive parts of speech, the syntactic function of which cannot be determined *a priori.* Thus the word *mothers,* for example, cannot be classed *a priori* either among the nouns or among the verbs and is thus incorporated into a new SUB/VER category possessing special characteristics.

- *Internally ambiguous part of speech:* A group of words belonging to an externally fully determined category but able to fulfill different functions within this category. (Ex. SHEEP = SUB - indeterminate number, HAVE = VER -indeterminate person and mood, etc.)

- *Part of speech with mixed ambiguity:* A group of words belong to an externally *and* internally ambiguous category (Ex. EQUAL = ADJ - gender determined by antecedent of following noun, and EQUAL = VER - indeterminate person and mood).

In the main body of the preceding paper fully determined parts of speech are designated by the symbol A_i while the ambiguous ones bear the symbol X_j.

B. First results

- *Words and occurrences:* The corpus examined contains 1,014 occurrences representing 341 distinct words. This proportion of 33.6 per cent of words seems sufficiently significant to us to be retained for programming purposes. It appears to occur consistently in other corpora of the same type and *size* (with minor ± variations of course.)

- *Sentences:* The corpus contains 48 sentences, of which:
eight are titles

one is interrogative
thirty nine are declarative

Their relative length is reflected in *Table I* below.

While sequences of one to 20 occurrences comprise one half of all the sentences examined, the number of occurrences contained in this group is only 26.72 per cent. In order to translate 86.65 per cent of occurrences one should be able to process sentences of at least 45 occurrences, and 100 per cent solution could only be achieved with sentences of at least 57 occurrences.

- *Parts of speech:* To study this problem we started out with a traditional syntactic analysis. The resulting *a posteriori* classification appears in columns A and B of *Table II.*

As a next step we sorted all the words alphabetically and proceeded with indexing them according to Webster's Collegiate Dictionary. Of the 341 words (1,014 occurrences) examined, 12(16) did not appear in the dictionary, 179(493) were fully determined, and 150 (505) were found to be externally ambiguous, that is their syntactic function could not be determined except *a posteriori* (See columns C-D, E-F, G-H of *Table II).*

Table II only takes account of external ambiguity. If we were also to take into account internal ambiguity (for the verb alone, for example), the number of words and occurrences determined *a priori* would fall to 151 and 420 respectively, and the degree of their determination would be reduced to 44 and 41.4 per cent (for the verb: 6 and 22.1 per cent). However, the verb is not the only part of speech afflicted with internal ambiguity. The same problem arises as far as pronouns, certain conjunctions and certain punctuation marks are concerned. All in all, the actual degree of *a priori* determination in the English language appears to be deceivingly small.

Table III contains a list of 149 ambiguous words and 448 ambiguous occurrences (the comma, that is: one word and 57 occurrences, will be the object of a special study). This table shows the way in which these external ambiguities have been resolved *a posteriori.*

The corpus examined has so far produced five degrees of external ambiguity (apart from the comma). This means that to determine the syntactic value of the 149 ambiguous words, we actually had to choose between 364 theoretically possible values as shown in *Table IV.*

Taking into account the limited nature of the text examined, the list in *Table III is* most certainly far from being exhaustive. On the other hand, the classification and the figures given took: no account of mixed and internal ambiguities. Their classification would introduce into this list a great many additional words, parts of speech and supplementary sub-categories.

- *Syntactic relations between the occurrences: Table 7* shows that the cases where several nouns (or pronouns) occur in a sentence together with an equally important or even larger number of occurrences identified as verbs are the rule and not the exception. *Table IV* indicates that the degree of determination of both is scarcely one half. *Table II* shows that the ambiguities SUB/VER are precisely the most numerous of all. For these reasons the problem of syntactic analysis in English is much more complicated than in Russian, German or even in French.

The sample examined is of course much too small to serve as a basis for more than tentative conclusions and it is not sufficient to yield universally valid statistics. However, the figures quoted are sufficient to indicate that, in translation, one *may* run into texts presenting at least the difficulties enumerated above. They are also sufficient to formulate certain hypotheses concerning the most profitable direction of future work.

(See attached tables)

TABLE I - LENGTH OF SENTENCES

A	B	C	D	E	F	G
Length in occurrences by groups of five	Number of sentences in each group	Number of sentences (cumulative)	Number of occurrences in each group	Number of occurrences (cumulative)	C expressed as a % of the total	E expressed as a % of the total
1 - 5	5	5	13	13	10,41	1,28
6 - 10	7	12	61	74	25,00	7,36
11 - 15	6	18	75	149	37,50	14,79
16 - 20	7	25	122	271	52,08	26,72
21 - 25	8	33	186	457	68,75	45,06
26 - 30	4	37	110	567	77,08	55,91
31 - 35	4	41	134	701	85,40	69,13
36 - 40	2	43	73	774	89,58	76,33
41 - 45	3	46	126	900	95,83	88,65
57 - 60	2	48	114	1,014	100	100
TOTAL	48	48	1,014	1,014		

TABLE II - PARTS OF SPEECH

Parts of speech		Total number (identified a posteriori by contextual analysis		Identified univocally a priori (on the basis of Webster)		Ambiguous in Webster. Identified a posteriori		Not included in Webster		Degree of a priori ** determination	
		A-Words	B-Occur	C-Words	D-Occur	E-Words	F-Occur	G-Words	H-Occur	I-Words	J-Occur
1	ADJ	78	126	43	67	35	59			55,1	53,0
2	ADV	28,5*	67	17	30	11,5*	37			59,6	44,8
3	ARD	1	80	00	00	1	80			00,0	00,0
4	ARI	1,5*	25	1	12	0,5*	13			66,7	40,0
5	CNJ	9,5*	60	3	18	6,5*	42			31,7	30,0
6	PRE	14	109	4	50	10	59			28,5	45,9
7	PRO	9,5*	50	3	26	6,5*	24			31,6	52,0
8	SUB	98	185	53	104	45	81			54,3	56,2
9	VER	65	159	32	106	33	53			46,1	60,0
10	CHI	14	14	14	14	00	00			100,0	100,0
11	SIP	10	123	9	66	1	57			10,0	46,3
12	NOP	10	14	0	00	00	00	10	14	00,0	00,0
13	FOR	2	2	0	00	00	00	2	2	100,0	100,0
	TOTAL	341	1,014	179	493	150	505	12	16	52,5	48,6

*) The notation "0.5 words" means that the same originally ambiguous word occurs in the text as two different unambiguous parts of speech. Thus, for instance, "If a car ...; If a and b ..."

**) I - C/A; J - D/B

TABLE III - PARTS OF SPEECH WITH EXTERNAL AMBIGUITY

CLASSIFICATION OF AMBIGUITIES AND THEIR RESOLUTION A POSTERIORI

		Words	ADJ	ADV	ARD	ARI	CNJ	INT	PRE	PRO	SUB	VER	TOT. WORDS	Occur	ADJ	ADV	ARD	ARI	CNJ	INT	PRE	PRO	SUB	VER	TOT. OCCUR.
I	1	ADJ/ADV	2	3,5									5.5		3	4									7
	2	ADJ/PRO	5							3			8		5							16			21
	3	ADJ/SUB	10,5								0,5		11		20							5			25
	4	ADJ/VER	4,5									4	8,5		12									5	17
	5	ADV/ARD			1								1				80								80
	6	ADV/CNJ		1									1			1									1
	7	ADV/PRE		1					4				5			5					23				28
	8	ADV/SUB		1									1			1									1
	9	ARI/SUB				0,5					0,5		1					13					4		17
	10	CNJ/PRE							1				1								4				4
	11	CNJ/PRO					1						1						1						1
	12	CNJ/SUB					2						2						18						18
	13	PRO/SUB								1			1									3			3
	14	SUB/VER									34	23	57										58	44	102
II	1	ADJ/ADV/CNJ					1						1						1						1
	2	ADJ/ADV/PRE							3				3								13				13
	3	ADJ/ADV/PRO	1							1			2		1							1			2
	4	ADJ/ADV/SUB	1								1		2		2								1		3
	5	ADJ/ADV/VER		1									1			8									8
	6	ADJ/SUB/VER	5								9	5	19		8								13	2	23
	7	ADV/CNJ/PRE					1						1						1						1
III	1	ADJ/ADV/PRE/SUB							1				1									5			5
	2	ADJ/ADV/CNJ/PRO					0,5			0.5			1						15			2			17
	3	ADJ/ADV/INT/SUB	1										1		1										1
	4	ADJ/ADV/PRE/VER							1				1								14				14
	5	ADJ/ADV/SUB/VER	4,5									0,5	5		6									1	7
	6	ADV/CNJ/INT/PRO		1									1			3									3
	7	ADV/CNJ/PRE/PRO		1									1			9									9
	8	ADV/CNJ/INT/SUB		1									1			3									3
IV	1	ADJ/ADV/CNJ/SUB/VER	0,5									0.5	1		1									1	2
	2	ADJ/ADV/INT/PRO/SUB								1			1									2			2
	3	ADJ/ADV/CNJ/PRE/SUB		0,5			0,5						1			2			2						4
V	1	ADJ/ADV/INT/PRE/SUB/VER		0.5			0,5						1			1			4						5

TABLE IV - POTENTIAL AND EFFECTIVE SYNTACTIC VALUES OF WORDS WITH EXTERNAL AMBIGUITY

		potential		effective
74	ADJ		35	
39,5	ADV		11,5	
1	ARD		1	
1	ARI		0,5	
13	CNJ		6,5	
5	INT		0	
15	PRE		10	
16	PRO		6,5	
106	SUB		45	
93,5	VER		33	
364			149	

TABLE V - SYNTACTIC RELATIONS

Sentence No.	Number of words	ADJ	ADV	ARD	ARI	CNJ	NOP	PRE	PRO	SUB	VER	CHI	FOR	,	.	(	)	-	;	"	?
1/30	21	4	1	2	1			4		6	2				1						
2/31	9	3		1		1				1	2				1						
3/32	11	1		2				2	1	2	2				1						
4/33	36	2	2	4	1	4		4	2	7	6			3	1						
5/34	27	1	7	1	1	3		3	1	4	5				1						
6/35	21	2	3	1	2	1		3	1	4	3				1						
7/36	11	3	2	1				1	1	1	1				1						
8/37	37	8	4	6		5				5	8				1						
9/38	30	4	1	6		1		5		7	3			2	1						
10/39	41	6	2	6		5			5	4	11				1				1		
11/40	17	4	3			1			1	2	4			1	1						
	261	38	25	30	5	21		22	12	43	47			6	11				1		

Para/C

ENGLISH CORPUS USED FOR A

PRELIMINARY EXAMINATION FOR SYNTAX

A. Probability the Basic Tool of Exterior System Design

Fundamental Notions -

In classical logic propositions are either true or false, with no room for doubt. We say "If it rains, I shall go to the movies, otherwise not." Uncertainty is indicated, but no measure of the uncertainty is available. My future action is undecided, depending on an event (the future state of the weather) of which I am in ignorance. The ignorance may be almost complete (if it rains the day Joe visits, whenever that is) or partial (if it rains an hour from now, which seems almost certain in view of the sky, the barometer, and the known position and movement of the front to the west). We should like to express the degree of credibility to be placed in the unqualified assertion, "I shall go to the movies", since the outcome "depends on chance."

To define chance, we assume an experiment, real or imaginary which may materialize in two or more outcomes which are possible under what are, in fact or to the best of our knowledge, identical conditions. Under these conditions, the result is said to be due to chance. It should be noted how subjective this definition is, depending as it does on the body of knowledge. Probability is the measure of chance.

A working definition of numerical probability is given below, but it is not completely satisfactory because it depends on the word likely which, of course, means probable. There is difficulty in defining probability, as there is in defining most fundamental concepts. Such units as time and length are quite difficult to define but nonetheless are used constantly by the engineer, and with considerable utility.

There is an extensive literature on the subject of the logical bases of probability. There are several approaches to this philosophical question. One, represented by Reichenbach[18] and Von Mises[19] attempts to define probability on a frequency basis; that is, if the number of experiments is allowed to approach infinity, then the probability of a favourable outcome may be defined as the limit of the proportion of the experiments which are favorable. A second approach, represented by Carnap,[20] Jeffreys [21] and Keynes[22] views probability as a logical relation analogous to that of logical implication but admitting of degrees. A third approach, re-

represented by Koopman[22] and Kolmogoroff,[17] attempts to define probability on an axiomatic basis; it states that probability is a game to be played according to certain rules, worked out on a strict mathematical basis. Our own use will be more like the last.

B 4-3 Total, Compound and Conditional Probability. Total probability is defined as the probability of any one of several mutually exclusive outcomes. It is equal to the sum of the individual probabilities, as we shall prove, although it is intuitively obvious. Consider an experiment with several possible different results, which we shall designate A, B and so forth, through K. Of the n outcomes which are equally possible, n_1 lead to A, n_2 to B, and so forth, with $\sum n_1 = n$

C Compound probability is defined as the probability of the joint occurrence of a pair of specified outcomes in two experiments. The two experiments may be identical or different. The outcome of one may be dependent on the other. Suppose, for example, that the experiments are the arrival of successive northbound automobiles at an intersection and the outcomes in which we are interested are whether or not the autos turn left. Now it frequently happens that a driver intends to turn left at some point along a thoroughfare and will do so at the first good opportunity. Such an opportunity is especially likely to present itself at a particular intersection if the car in front turns left. Hence the successive experiments are not independent of one another. The conditional probability that the nth car will turn left if the (n - l)st car has turned left is higher than the conditional probability that the nth car will turn left if the (n - l)st car has not. The unconditional probability that the nth car will turn left, in the absence of knowledge of the behaviour of the previous car, is intermediate between the two. Now we can consider the conditional probability that the second car will turn if the first one has; we can also consider the conditional probability that the first car had turned if we know that the second one did. Because these two conditional probabilities may not be equal, we must consider the experiments sequentially.

D 4-4. Markoff Chains. A sequence of events in which the probability of a particular outcome on the nth event is fully determined by the outcome of the (n - l)st event is called a Markoff process or Markoff chain. We shall take the automobiles turning left at an intersection (Sec. 4-3) as an example of a Markoff chain. Assume that the probability of an automobile turning left is 0.2 if the previous auto turned left and 0.1 if the previous auto did not. What is the probability for the nth auto turning? We know that the probability P(k) that the kth car turns, plus the probability that it does not turn, is equal to one. Hence, we can write

P(k + 1) in terms of P(k):

$$P(k+1) = 0.2\,P(k) + 0.l\,[l - P(k)] = 0.1 + 0.1\,P(k) \quad (4\text{-:}$$

Let the probability for the zeroth car be P(0), an unknown.